

## HIERARCHICAL BAYES ESTIMATION OF RESPONSE RATES FOR AN EXPENDITURE SURVEY

Yong You and Philip Reiss<sup>1</sup>

### ABSTRACT

Statistics Canada's Homeowner Repair and Renovation Survey (HRRS) is an annual supplement to the Canadian Labour Force Survey (LFS), measuring expenditures for additions, renovations, repairs and maintenance, and installation and replacement of household equipment. Currently the HRRS nonresponse is adjusted for by the same method as is used for LFS nonresponse: the sample is divided into response homogeneity groups (RHG), and weights within each RHG are multiplied by the inverse of its weighted response rate. This method produces somewhat unstable HRRS nonresponse adjustment factors, due to the small size of many of the RHG's. In this paper, a compound beta-binomial model is presented to smooth the direct survey estimate of the response rate of HRRS. A hierarchical Bayes approach and the Gibbs sampling method are employed to obtain the posterior estimate of the response rate, and the uncertainty of the estimator is measured by its posterior variance.

KEY WORDS: Beta-binomial; Gibbs sampling; Hierarchical Bayes; Response rate; Small area; Smoothing.

### RÉSUMÉ

L'Enquête sur les réparations et les rénovations effectuées par les propriétaires-occupants (ERRP) de Statistique Canada est un supplément à l'Enquête sur la population active (EPA) qui mesure les dépenses pour les additions, les rénovations, les réparations et l'entretien, et l'installation et le remplacement de l'équipement ménager. Présentement, on utilise la même méthode d'ajustement pour la non-réponse de l'ERRP que pour la non-réponse de l'EPA: l'échantillon est divisé en groupes de réponse homogènes (GRH's), et les poids dans chaque GRH sont multipliés par l'inverse de son taux de réponse pondéré. Cette méthode produit des facteurs d'ajustement quelque peu instables à cause de la petite taille de beaucoup de GRH's. Dans cet article, un modèle bêta-binomial composé est utilisé pour lisser les estimations directes du taux de réponse de l'ERRP. Une approche hiérarchique de Bayes et la méthode d'échantillonnage de Gibbs sont utilisées pour obtenir l'estimation a posteriori du taux de réponse, et l'incertitude de l'estimateur est mesurée par la variance a posteriori.

MOTS-CLÉS: Bêta-binomial; échantillonnage de Gibbs; hiérarchique de Bayes; taux de réponse; petites régions; lissage.

### 1. INTRODUCTION

Statistics Canada's Homeowner Repair and Renovation Survey (HRRS) measures levels of expenditure by Canadian homeowners for additions, renovations, repairs and maintenance, and installation and replacement of household equipment. The HRRS is conducted each March as a supplementary survey to the Canadian Labour Force Survey (LFS). The monthly LFS follows a rotating panel design, with households remaining in the sample for six consecutive months (Gambino et al., 1998). Respondents in four of the March LFS sample's six rotation groups are asked the HRRS questions after completing the LFS (Statistics Canada, 1998). As a supplement to LFS, HRRS has two types of non-response: non-response to LFS (usually around 5%), and

non-response to HRRS itself by LFS respondents (about 12% of the latter). The second type of non-response—more precisely, the need to increase the weights to account for it—motivated the present study.

Currently, HRRS non-response is adjusted for by the same method as is used for LFS non-response. The sample is divided into response homogeneity groups (RHG's) within each province. In most cases, an RHG is the intersection of (i) an Employment Insurance Economic Region, (ii) an LFS sample design type (e.g., urban cluster design or rural three-stage design), and (iii) a rotation group. Within each RHG, weights are multiplied by the inverse of the weighted response rate. This method produces somewhat unstable HRRS non-response adjustment, due to the small size of many

<sup>1</sup> Yong You and Philip Reiss, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6.

of the RHG's. To remedy this instability, in this paper we propose a hierarchical Bayes model to derive a model-based estimator of the response rate for each RHG. Such a model allows RHG's to borrow strength from related regions to smooth the empirical estimates of response rates, and thereby to improve the accuracy of the estimates.

Some recent work has considered the Bayesian analysis of binary data in small area estimation. Datta et al. (1999) obtained hierarchical Bayes estimates of unemployment rates for the U.S. states using a cross-sectional and time series model. Maiti (1998) proposed a hierarchical Bayes model involving the Poisson distribution and a log-linear model to smooth the relative risks of a disease for small areas. He and Sun (1998) used a beta-binomial model to obtain the posterior success rates for small areas. Rosner (1989) presented multi-variate methods for clustered binary data with more than one level of nesting. In particular, he presented a compound beta-binomial model which, in the absence of covariates, generalizes the beta-binomial distribution to more than one level of nesting. In this paper, such a compound beta-binomial model without covariates will be fit to the HRRS response rates, which we treat as direct survey estimates of response probability. A hierarchical Bayes approach is presented for smoothing these direct survey estimates. We use the Gibbs sampling method to obtain the posterior estimate of the response rate, and a measure of uncertainty of the estimate is provided by the posterior variance.

In principle, the status of HRRS as a supplement to LFS allows us to use any LFS variable as an additional predictor of non-response. However, logistic regression of response status on several such variables (such as household size, dwelling type and level of education) found none that were significantly associated with response probability. On the other hand, each of the three variables used to define the RHG's (see above) was a very significant predictor of response status.

## 2. HIERARCHICAL BETA-BINOMIAL MODEL

Let  $p$  denote the national level response rate, let  $p_i$  denote the province-level response rate for the  $i$ -th province, and let  $p_{ij}$  denote the response rate of the  $j$ -th RHG in the  $i$ -th province. Furthermore, let  $n_{ij}$  be the sample size of the  $ij$ -th RHG, and  $y_{ij}$  the number of respondents among the  $n_{ij}$  samples. We consider the following hierarchical Bayes compound beta-binomial model:

### Response model:

$$y_{ij}|p_{ij} \sim \text{Binomial}(n_{ij}, p_{ij}), \quad i=1, \dots, m; \quad j=1, \dots, m_i.$$

### Population model:

$$p_{ij} \sim \text{Beta}(\lambda_1 p_i, \lambda_1 (1-p_i)), \quad i=1, \dots, m; \quad j=1, \dots, m_i,$$

$$p_i \sim \text{Beta}(\lambda_0 p, \lambda_0 (1-p)), \quad i=1, \dots, m.$$

### Prior distributions:

$$\lambda_1 \text{ Gamma}(a_1, b_1), \quad \lambda_0 \text{ Gamma}(a_0, b_0), \quad a_1 > 0, \quad b_1 > 0, \quad a_0 > 0, \quad b_0 > 0.$$

In the above model,  $m$  is the number of provinces;  $m_i$  is the number of RHG's within the  $i$ -th province;  $\lambda_1$  and  $\lambda_0$  are hyperparameters; and  $a_1, b_1, a_0, b_0$  are known constants.

For each RHG, the number of respondents  $y_{ij}$  is modelled as a random variable with a binomial distribution  $\text{Binomial}(n_{ij}, p_{ij})$ . The  $y_{ij}$ 's are independent of each other. The direct survey estimate of the response rate  $p_{ij}$  is given by

$$\hat{p}_{ij} = \frac{y_{ij}}{n_{ij}}. \quad (1)$$

The standard error of  $\hat{p}_{ij}$  is estimated by

$$\hat{s}_{ij} = \sqrt{\frac{\hat{p}_{ij}(1-\hat{p}_{ij})}{n_{ij}}}. \quad (2)$$

Note that  $\hat{s}_{ij}$  is defined only for  $0 < \hat{p}_{ij} < 1$ . For some RHG's, the sample size  $n_{ij}$  is very small; the direct estimate  $\hat{p}_{ij}$  is therefore not reliable for these groups. We would like to obtain the posterior estimate of  $p_{ij}$  to smooth the direct estimate  $\hat{p}_{ij}$ . That is, given  $Y = (y_{ij}; i=1, \dots, m, j=1, \dots, m_i)$ , we are interested in estimating  $E(p_{ij}|Y)$  and the posterior variance  $V(p_{ij}|Y)$ .

In fact, the compound beta-binomial model makes possible two levels of smoothing: at the RHG level and at the province level. We can estimate  $E(p_{ij}|Y)$  and  $E(p_i|Y)$  at the same time. The direct estimate of  $p_i$  at the province level is given by

$$\hat{p}_i = \frac{\sum_{j=1}^{m_i} y_{ij}}{\sum_{j=1}^{m_i} n_{ij}}. \quad (3)$$

However, since the provinces of Canada vary greatly in size and population structure and  $\hat{p}_i$ 's tend to fall as we move from east to west,  $\hat{p}_i$  can be smoothed by  $E(p_i|Y)$  through the compound beta-binomial model, although our main interest is to smooth the  $\hat{p}_{ij}$ 's for the RHG's within provinces.

The two levels of smoothing can be better understood as follows:

(i) Conditional on  $p_i$ , the distribution of  $p_{ij}$  is modelled by a beta distribution with parameters  $\lambda_i p_i$  and  $\lambda_i(1-p_i)$  for some  $\lambda_i > 0$ . Therefore, we have  $E(p_{ij} | \lambda_i, p_i) = p_i$ . Thus we assume within the  $i$ -th province, the response rates  $p_{ij}$  are random variables with conditional mean  $p_i$ , the province-level response rate.

(ii) Conditional on  $p$ , the distribution of  $p_i$  among provinces is modelled by a beta distribution with parameters  $\lambda_0 p$  and  $\lambda_0(1-p)$  for some  $\lambda_0$ . Then we have  $E(p_i | \lambda_0, p) = p$ . Thus the provincial response rate  $p_i$  is a random variable with mean equal to the national response rate  $p$ . The national response rate  $p$  can be calculated as

$$p = \frac{\sum_{i=1}^m \sum_{j=1}^{m_i} y_{ij}}{\sum_{i=1}^m \sum_{j=1}^{m_i} n_{ij}}. \quad (4)$$

Since calculation of  $p$  is based on the overall sample sizes and overall number of respondents for the whole nation,  $p$  is fairly stable at the national level. Thus  $p$  is considered known in the model.

### 3. GIBBS SAMPLING INFERENCE

We are interested in estimating the posterior distributions of  $p_{ij}$  and  $p_i$ , in particular, the posterior means and posterior variances of  $p_{ij}$  and  $p_i$ . Direct evaluation involves high-dimensional integration and is not computationally feasible. Instead, we use the Gibbs sampling method. The full conditional distributions for the Gibbs sampler are given as follows:

$$(i) [p_{ij} | Y, p_i, p, \lambda_1, \lambda_0] \propto p_{ij}^{y_{ij} + \lambda_1 p_i - 1} (1 - p_{ij})^{n_{ij} - y_{ij} + \lambda_1(1-p_i) - 1};$$

$$(ii) [p_i | Y, p_{ij}, p, \lambda_1, \lambda_0] \propto$$

$$\frac{(\prod_{j=1}^{m_i} p_{ij})^{\lambda_1 p_i} (\prod_{j=1}^{m_i} (1 - p_{ij}))^{\lambda_1(1-p_i)}}{[\Gamma(\lambda_1 p_i) \Gamma(\lambda_1(1-p_i))]^{m_i}} p_i^{\lambda_0 p} (1 - p_i)^{\lambda_0(1-p)};$$

$$(iii) [\lambda_1 | Y, p_{ij}, p_i, p, \lambda_0] \propto$$

$$\prod_{i=1}^m \left[ \frac{\Gamma(\lambda_i)}{\Gamma(\lambda_i p_i) \Gamma(\lambda_i(1-p_i))} \right]^{m_i} \prod_{i=1}^m (\prod_{j=1}^{m_i} p_{ij})^{p_i} (\prod_{j=1}^{m_i} (1 - p_{ij}))^{1-p_i} \lambda_i^{a_1 - 1} e^{-b_1 \lambda_i};$$

$$(iv) [\lambda_0 | Y, p_{ij}, p_i, p, \lambda_1] \propto$$

$$\left[ \frac{\Gamma(\lambda_0)}{\Gamma(\lambda_0 p) \Gamma(\lambda_0(1-p))} \right]^m \left[ \prod_{i=1}^m p_i^{p_i} (1 - p_i)^{1-p_i} \right]^{\lambda_0} \lambda_0^{a_0 - 1} e^{-b_0 \lambda_0}.$$

The full conditional distribution of  $p_{ij}$  is a beta distribution with parameters  $y_{ij} + \lambda_1 p_i$  and  $n_{ij} - y_{ij} + \lambda_1(1 - p_i)$ ; thus updating  $p_{ij}$  is simple. However, the Metropolis-Hastings algorithm is needed to generate samples from the conditional distributions of  $p_i$ ,  $\lambda_1$  and  $\lambda_0$ . Since the conditional distribution of  $p_{ij}$  has a closed form, the Rao-Blackwellized estimators (Gelfand and Smith, 1990, 1991) of  $E(p_{ij} | Y)$  and  $V(p_{ij} | Y)$  can be constructed. Suppose a sample of size  $G$  is generated from the Gibbs sampler. Noting that

$$E(p_{ij} | Y, p_i, p, \lambda_1, \lambda_0) = \frac{y_{ij} + \lambda_1 p_i}{n_{ij} + \lambda_1} \quad (5)$$

and

$$V(p_{ij} | Y, p_i, p, \lambda_1, \lambda_0) = \frac{(y_{ij} + \lambda_1 p_i)(n_{ij} - y_{ij} + \lambda_1(1 - p_i))}{(n_{ij} + \lambda_1)^2 (n_{ij} + \lambda_1 + 1)}, \quad (6)$$

it follows from  $E(p_{ij} | Y) = E[E(p_{ij} | Y, p_i, p, \lambda_1, \lambda_0) | Y]$  and equation (5) that the Rao-Blackwellized estimator of the posterior mean of  $p_{ij}$  is given by

$$\hat{p}_{ij}^{RB} = \frac{1}{G} \sum_{k=1}^G \frac{y_{ij} + \lambda_1^{(k)} p_i^{(k)}}{n_{ij} + \lambda_1^{(k)}}. \quad (7)$$

To estimate the posterior variance of  $p_{ij}$ , noting that

$$V(p_{ij} | Y) = E(V(p_{ij} | Y, p_i, p, \lambda_1, \lambda_0)) + V(E(p_{ij} | Y, p_i, p, \lambda_1, \lambda_0)), \quad (8)$$

and using equations (5) and (6), we obtained the Rao-Blackwellized estimator of  $V(p_{ij} | Y)$  as

$$\hat{V}_{p_{ij}}^{RB} = \frac{1}{G} \sum_{k=1}^G \frac{(y_{ij} + \lambda_1^{(k)} p_i^{(k)})(n_{ij} - y_{ij} + \lambda_1^{(k)}(1 - p_i^{(k)}))}{(n_{ij} + \lambda_1^{(k)})^2 (n_{ij} + \lambda_1^{(k)} + 1)} + \frac{1}{G} \sum_{k=1}^G \frac{(y_{ij} + \lambda_1^{(k)} p_i^{(k)})^2}{(n_{ij} + \lambda_1^{(k)})^2} - \left( \frac{1}{G} \sum_{k=1}^G \frac{y_{ij} + \lambda_1^{(k)} p_i^{(k)}}{n_{ij} + \lambda_1^{(k)}} \right)^2. \quad (9)$$

The Rao-Blackwellized estimators can substantially reduce the simulation random errors (Gelfand and Smith, 1991; You and Rao, 1999). Thus the beta-binomial model enables us to find stable and closed form estimators for the posterior mean and posterior variance of  $p_{ij}$ .

#### 4. DATA ANALYSIS

In our data analysis, we use 1997 HRRS data, for which there are 538 RHG's across Canada. (Nine high-income RHG's were excluded from our study, and another eight were collapsed into other RHG's because their response rate were less than 50%.) The number of RHG's varies widely among the provinces, from 8 in Prince Edward Island to 153 in Ontario. Sample size within RHG's also differs substantially, varying from 1 to 201. Our goal is to smooth the direct survey estimate of the response rate, particularly for the RHG's with small sample size. Using (4), we calculated the national response rate as  $p=0.891$ . Prior distributions for the hyperparameters  $\lambda_1$  and  $\lambda_0$  are set as  $\text{Gamma}(1,0.001)$ , which is equivalent to an exponential distribution with mean 1000. We fit the compound beta-binomial model via Gibbs sampling using the BUGS 0.6 program (Spiegelhalter, Thomas, Best and Gilks, 1997).

Within each province, the HB estimator  $\hat{p}_{ij}^{RB}$  smoothed the direct estimator  $\hat{p}_{ij}$ , especially for the  $\hat{p}_{ij}$  with extreme large and small values. Figure 1 displays the direct survey estimates and HB estimates of the response rates of the RHG's in Newfoundland and P.E.I. and the corresponding standard errors. It is clear from Figure 1 that the HB method has smoothed the direct estimates, and moved the RHG level response rate toward the posterior province-level response rate. There is a substantial reduction in the standard errors, particularly for the RHG's with small sample sizes. For example, for RHG 17 in Newfoundland, the sample size is only 4 and the direct estimate  $\hat{p}_{ij}=0.75$  with standard error equal to 0.2165; whereas the Bayes estimate  $\hat{p}_{ij}^{RB}=0.898$  with standard error equal to 0.0302. For the RHG's with  $y_{ij}=n_{ij}$ , we get  $\hat{p}_{ij}=1$ . Thus we cannot use (2) to calculate the standard errors of these  $\hat{p}_{ij}$ . For these RHG's, the sample size  $n_{ij}$  is typically very small. However, by using the HB approach, we can obtain the smoothed estimate  $\hat{p}_{ij}^{RB}$  and the corresponding posterior variance as the measure of uncertainty associated with the  $\hat{p}_{ij}^{RB}$ . Overall, the HB method improves on the direct estimates at the RHG level considerably. Detailed results and

results for the other provinces can be obtained from the authors.

At the province level, Figure 2 shows the direct estimates and the HB estimates of the provincial level response rates and the national response rate. We note from Figure 2 that the provincial response rates have consistently moved toward the national response rate of 89.1%; provinces close to this national rate, such as Quebec, Ontario and Manitoba move very little, while extreme provinces such as Newfoundland and P.E.I. move the most.

#### 5. REMARKS AND FUTURE WORK

In this paper, we have presented a hierarchical Bayes compound beta-binomial model to smooth the direct survey estimates of the response rates of the RHG's in the HRRS. The proposed compound beta-binomial model smoothes both the RHG level and province level response rates at the same time. Also, the beta-binomial model leads to closed-form full conditional distributions for  $p_{ij}$ , which enables us to obtain the Rao-Blackwellized estimators for the posterior mean and posterior variance of  $p_{ij}$ . If auxiliary variables related to the response rate are available at the RHG level, probit or logistic random effect regression models (Ghosh et al., 1998) may be used to find the posterior estimates of  $p_{ij}$ . However, with the probit or logistic model, we cannot get closed-form full conditionals for  $p_{ij}$ . Nevertheless, it may be interesting to compare the results under different models and make appropriate model comparisons.

At the present time, simulation studies are ongoing to determine whether the hierarchical Bayes non-response adjustment would significantly improve HRRS estimates. While such studies are difficult to make precise, it appears the reductions in MSE due to non-response are too slight to justify implementing the method for this survey. There are indications the method may be more helpful for surveys with higher non-response.

Figure 1: Estimates and Standard Errors (Newfoundland and P.E.I.)

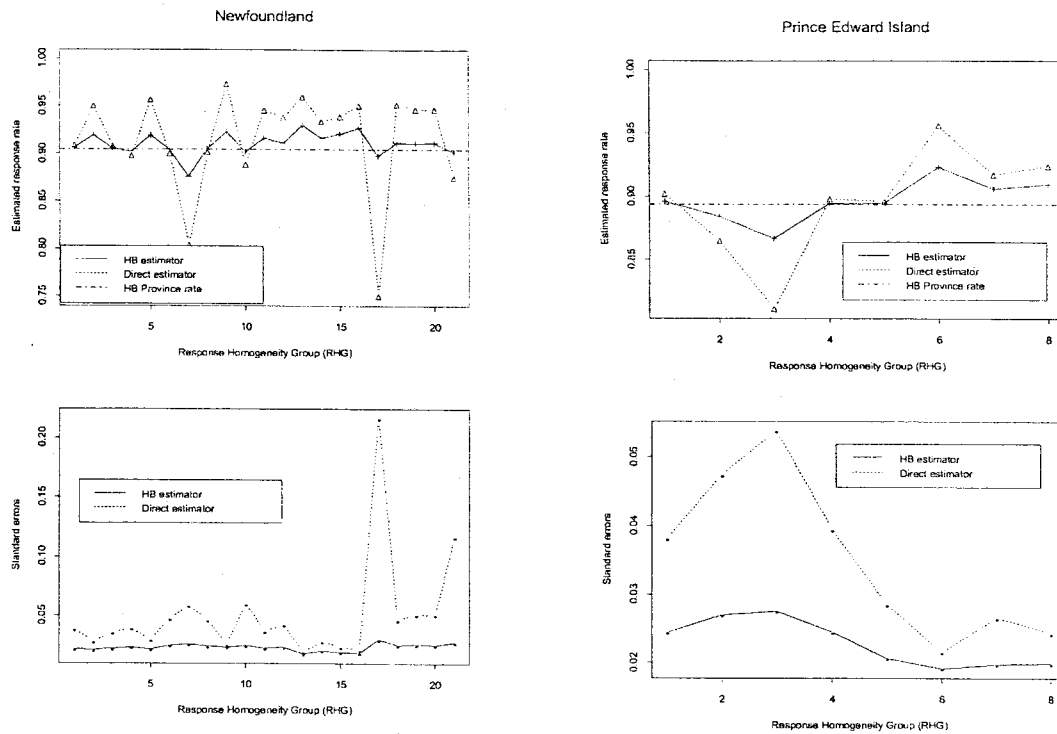
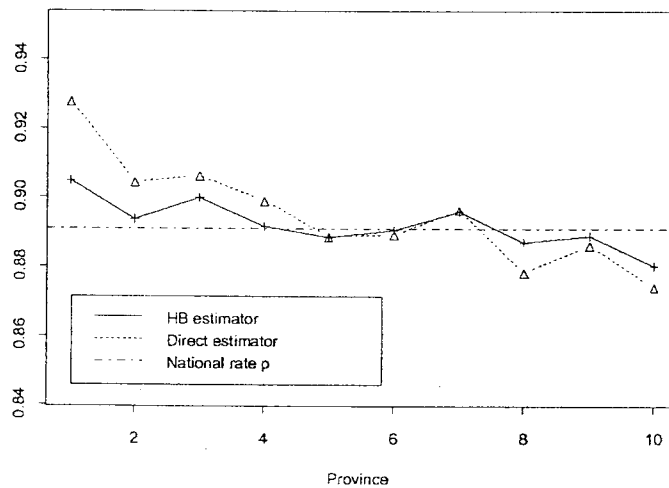


Figure 2: Estimated Province Level Response Rates



## ACKNOWLEDGEMENT

The authors gratefully acknowledge Professor J.N.K. Rao of Carleton University and Diane Stukel of Statistics Canada for their helpful suggestions

## REFERENCES

- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999) Hierarchical Bayes estimation of unemployment rates for the U.S. states. Manuscript.
- Dick, P. and You, Y. (1997) A hierarchical Bayes analysis of Census undercoverage. *Proceedings of Symposium 97: New Directions in Surveys and Censuses*, 101-105, Statistics Canada, Ottawa.
- Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B., and Lindeyer, J. (1998) *Methodology of the Canadian Labour Force Survey*. Catalogue no. 71-526-XPB, Statistics Canada, Ottawa.
- Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Gelfand, A.E. and Smith, A.F.M. (1991) Gibbs sampling for marginal posterior expectations. *Communications In Statistics: Theory and Methods*, 20, 1747-1766.
- Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998) Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.
- Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal (with discussion). *Statistical Science*, 9, 55-93.
- He, Z. and Sun, D. (1998) Bayes estimation of success rate for small areas. Manuscript.
- Maiti, T. (1998) Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Rosner, B. (1989) Multivariate methods for clustered binary data with more than one level of nesting. *Journal of the American Statistical Association*, 84, 373-380.
- Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1997) *BUGS 0.6: Bayesian inference Using Gibbs Sampling (Addendum to Manual)*. MRC Biostatistics Unit, Cambridge.
- Statistics Canada (1998). *Homeowner Repair and Renovation Expenditure 1996*. Catalogue no. 62-201-XPB.
- You, Y and Rao, J.N.K. (1997) Hierarchical Bayes small area estimation using multi-level models. *Proceedings of Symposium 97: New Directions in Surveys and Censuses*, 278-281, Statistics Canada, Ottawa.
- You, Y and Rao, J.N.K. (1999) Hierarchical Bayes estimation of small area means using multi-level models. Invited paper, IASS Satellite Conference on Small Area Estimation, Riga, Latvia.